Increasing Object-Level Reconstruction Quality in Single-Image 3D Scene Reconstruction

Anna Ribic Antonio Oroz Meikel Kokowski Franz Srambical
Technical University of Munich
{firstname}.{lastname}@tum.de

Abstract

1. Introduction

While humans can easily infer the 3D structure as well as the complete (panoptic) semantics of a scene from a single image, this task has been a longstanding challenge in the field of computer vision. The task fundamentally prerequisites learning a strong prior of the 3D world. Traditional methods have made significant strides, from generating geometrically coherent structures [? ?] to learning different instance semantics [? ? ?]. More recent approaches directly learn the 3D panoptic semantics as a whole [? ?], yet they fall short in capturing the intricate details and nuances at the object level. This paper introduces a novel approach to bridge this gap by integrating a specialized object-level model into the reconstruction process, thereby leveraging the specialized model's object-priors.

Our approach models panoptic 3D reconstruction as a two-stage problem. We first use the model of ?] to create an initial reconstruction. Then, we leverage the instance masks to extract the object geometries out of the reconstructed scene. We input each of the extracted objects along with cropped images from the scene and text labels into a diffusion model [?] to refine the rough object-level geometries. Finally, we integrate the refined object geometries back into the initial scene reconstruction to obtain a complete and refined panoptic 3D scene reconstruction.

In summary, our main contributions are as follows:

- We propose a novel approach to panoptic 3D reconstruction involving an inference pipeline that leverages object-level reconstruction models to refine the output of a 3D scene reconstruction backbone.
- We qualitatively demonstrate the effectiveness of our approach on the 3D-Front [?] dataset, showing significant improvements over the state-of-the-art.
- We show that fine-tuning SDFusion [?] on the input scene's object distribution (in our case the 3D-Future dataset [?]) significantly improves the quality of the refined objects.
- We propose *weighted masking*, a novel technique to integrate masking uncertainty into the object-level reconstruction process.
- We introduce a conceptually simple yet effective method for shape alignment, which outperforms rigid alignment methods in our experiments.
- We openly release our model code, training and inference pipelines, as well as our newly constructed variation of the 3D-Front dataset to facilitate future research in the field.

2. Related Work

2D panoptic segmentation 2D panoptic segmentation merges semantic and instance segmentation, providing detailed pixel-level parsing of images, capturing both general categories (semantic segmentation) and individual object identities (instance segmentation) [?]. Since the original task formulation by ?], a number of works have been proposed to solve the task [? ? ? ? ? ? ? ? ? ? ? ? ?], while more recent approaches [?] try to unify image segmentation in its entirety.

Single-view 3D reconstruction The work by ?] was the first notable attempt at reconstructing 3D scenes from unordered photo collections. Since then, the field of image-based 3D reconstruction has seen a number of advancements, culminating in the task of single-view 3D reconstruction [? ? ? ? ? ? ?].

Shape priors ?] note that the task of single-view 3D reconstruction is non-deterministic, as there are many 3D shapes that can explain a given single-view input, and propose to use shape priors to shape the solution space such that the reconstructed shapes are realistic, but not necessarily the ground truth.

3D scene understanding The task of 3D scene understanding and panoptic reconstruction is analogous to its 2D counterpart and aims to infer the 3D structure and semantics of a scene, including the 3D layout, object instances, and their 3D shapes from images [**?**] or noisy geometry [**? ?**]. **?**] propose a method – henceforth called *Panoptic 3D* – to jointly solve the tasks of 3D scene understanding and single-view 3D reconstruction by lifting features produced by a 2D backbone into a 3D volume of the camera frustrum, and jointly optimizing for geometric reconstruction as well as semantic and instance segmentation.

Modality-conditioned shape generation 3D generative models represent objects in a variety of modalities, including point clouds [**? ?**], occupancy grids [**?**], meshes [**?**], and signed distance functions [**?**]. Furthermore, these models can also be distinguished by the type of input they take, such as incomplete shapes [**?**], images [**?**], text [**? ?**], or other modalities [**?**]. Notably, **?**] propose *SDFusion*, a 3D object reconstruction method conditioned on images, text and geometrical input.

Datasets Notable datasets in the field of panoptic 3D reconstruction include ScanNet [**?**] and Replica [**?**], which provide rich annotations for scene understanding tasks. Another such dataset, 3D-Front [**?**], provides comprehensive coverage of indoor scenes while offering detailed geometric reconstructions as well as semantic and instance segmentation annotations. The synthetic 3D dataset contains 6,801 mid-size apartments with 18,797 rooms populated by 3D shapes from the 3D-Future [**?**] dataset. The dataset's high-quality data acquisition process ensures accurate representations, establishing it as a valuable resource for advancing research in 3D panoptic reconstruction.

3. Method

We leverage Panoptic 3D [**?**] to predict the camera frustum geometry and associated 3D semantic and instance labels within the image. This model provides us with both 2D and 3D representations of detected objects. For each detected object, we use the 2D instance mask to extract RGB crops of the input image, and the 3D instance mask to extract the corresponding 3D geometry. The extracted geometry along with the semantic label are subsequently input into the SDFusion model for shape reconstruction.

SDFusion employs task-specific encoders ([**? ?**]) to process the 2D image and text embeddings, while simultaneously embedding the 3D shape into a latent space using a pre-trained vector quantized variational autoencoder (VQ-VAE) [**?**]. Noise is then introduced to the shape latent via forward diffusion, which is followed by a concatenation of the conditional embeddings. This serves as input to the 3D U-Net [**?**] denoising network which reconstructs the latent code. Within the denoising U-Net, cross-attention is applied along the concatenated latent code to modulate the denoising process. Ultimately, the VQ-VAE decoder reconstructs the shape.

The output of SDFusion is front-facing and might not align with the object's orientation in the reconstructed 3D scene. To address this, we employ a registration algorithm to ensure proper alignment of the reconstructed objects within the scene.

Panoptic 3D Scene Reconstruction Panoptic [**?**] utilizes a single RGB image as input to simultaneously reconstruct the scene geometry and predict both semantic and instance segmentation labels for the reconstructed scene. To achieve this, Panoptic employs a ResNet-18 encoder for feature extraction from the input image. Subsequently, these features are utilized to predict both a 2D depth map and 2D instance mask through a depth encoder and a Mask R-CNN applied directly to the ResNet-18 features. The depth map facilitates the backprojection of features into a sparse volumetric grid, and the 2D instance mask is propagated to serve as a seed for the 3D instance mask prediction. Finally, a U-Net architecture processes the sparse backprojection to forecast occupancy, distance field, and both semantic and instance labels for each individual occupancy within the grid.

SDFusion SDFusion [**?**] utilized a signed distance field as its primary input. Additionally, it leverages an RGB image and a textual representation as conditional inputs to guide the reconstruction process. A variational autoencoder compresses the signed distance field into a latent space representation. Within the latent space, a diffusion process is implemented, gradually diffuses the latent code. Prior the denoising step, the conditional inputs are encoded and concatenated with the noisy latent code. A 3D U-Net architecture is then employed to denoise the latent code. Notably, an attention mechanism is incorporated, allowing the denoising network to selectively focus on relevant information from the conditional inputs during the denoising process. Finally, the decoder of the VAE reconstructs the SDF from the refined latent code, resulting in a reconstructed 3D shape.

Registration To ensure proper integration if the reconstructed shapes within the 3D scene, we employ a registration process that aligns the initially front-facing SDFusion outputs. This process consists of 3 key steps:

1. **Floor Alignment:** To establish a common reference frame and facilitate subsequent orientation, we align the reconstructed object with the floor plane of the 3D scene.
2. **Rotational Optimization:** Following floor alignment, the object is systematically rotated through 16 discrete, uniformly distributed positions around its y-axis. This exploration covers a diverse range of potential orientations.

3. **Selection based on Similarity:** The final step involves selecting the orientation that minimizes the per-point difference between the reconstructed object and the corresponding elements within the scene. This metric serves as a quantitative measure of alignment accuracy.

4. Results

Our study demonstrates the efficacy of our methodology in enhancing the visual presentation of objects within 3D reconstructions. As depicted in Figure 1, the unprocessed Panoptic reconstruction exhibits visual artificats and irregularities, while our approach yields smoothed surfaces, facilitating improved recognition through visual observation. Additionally, we observe the sensitivity of our alignment procedure to the instance masks generated by Panoptic. Although the predicted objects maintain smoothness (likely due to the conditioning in SDFusion), our alignment algorithm occasionally results in object intersection, as illustrated in Figure 2.

5. Conclusion, Limitations & Future Work

In summary, our method has demonstrated the efficacy of employing a reconstruction combined with a diffusion model to enhance the aesthetic quality of 3D instances in a reconstructed scene. Furthermore, our approach enables a simple process to create virtual environments from single images, which could be utilized in games, and virtual or augmented reality settings. Utilizing a multi-modal diffusion model instead of leveraging high-quality 3D objects directly (such as CADs), leads to another advantage, as the inputs can be customized enabling fine-grained control of the reconstructed scene.

While our results are promising, we also point out a few limitations to our method. The first one concerns the detection of objects. While the panoptic reconstruction model doesn't necessarily need to detect an instance to reconstruct its approximate shape, SDFusion is only applied to detected objects. Therefore undetected objects would not be refined, or if they are misidentified as a part of another instance, might be lost entirely. These noisy instance segmentations raise another issue as well, as our scale and position estimates rely on the instance panoptic properties. Therefore instance segmentations which include parts of other instances or parts of the wall/floor, lead to larger, misplaced refined instances.

To improve upon the mentioned limitations we have two directions in which future work could follow up. The first one could implement end-to-end training with adapted losses which would punish misidentified instances more. The second improvement would focus on the merging process. A pose estimation network, which predicts the scale and rotation of the instance might alleviate our issues arising from the usage of the chamfer distance and bounding boxes for these two properties, as both are vulnerable to outliers. Another area of interest could be themed/edited instance reconstructions through descriptions that can be added during inference, as the SDFusion paper shows interesting results based on detailed object descriptions.